

Statistical properties of damped Lyman-alpha systems from Sloan Digital Sky Survey DR12

Simeon Bird,^{1*} Roman Garnett,² Shirley Ho^{3 4 5}

¹*Department of Physics & Astronomy, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA*

²*Department of Computer Science and Engineering, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130, USA*

³*Department of Physics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA*

⁴*Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA*

⁵*Department of Physics, University of California, Berkeley, CA 94720*

20 December 2016

ABSTRACT

We present new estimates for the statistical properties of damped Lyman- α absorbers (DLAs). We compute the column density distribution function at $z > 2$, the line density, dN/dX , and the neutral hydrogen density, Ω_{DLA} . Our estimates are derived from the DLA catalogue of Garnett et al. (2016), which uses the SDSS–III DR12 quasar spectroscopic survey. This catalogue provides a probability that a given spectrum contains a DLA. It allows us to use even the noisiest data without biasing our results and thus substantially increases our sample size. We measure a non-zero column density distribution function at 95% confidence for all column densities $N_{\text{HI}} < 5 \times 10^{22} \text{ cm}^{-2}$. We make the first measurements from SDSS of dN/dX and Ω_{DLA} at $z > 4$. We show that our results are insensitive to the signal-to-noise ratio of the spectra, but that there is a residual dependence on quasar redshift for $z < 2.5$, which may be due to remaining systematics in our analysis.

Key words: Galaxies: intergalactic medium. Quasars: absorption lines

1 INTRODUCTION

Damped Lyman- α absorbers (DLAs) are absorption systems corresponding to neutral hydrogen column densities $> 2 \times 10^{20} \text{ cm}^{-2}$ (Wolfe et al. 1986). Observed primarily in quasar spectra at $z > 2$, they arise from gas dense enough to self-shield from the effect of the ultra-violet background (UVB) (Cen 2012), but diffuse enough to have a low star formation rate (Fumagalli et al. 2015). DLAs dominate the neutral hydrogen budget of the Universe after reionisation (Gardner et al. 1997; Noterdaeme et al. 2012; Zafar et al. 2013; Crighton et al. 2015). Simulations indicate that they are connected with a wide variety of halo masses, peaking in the range 10^{10} – $10^{11} M_{\odot}$ (Haehnelt et al. 1998; Prochaska & Wolfe 1997; Pontzen et al. 2008). The DLA halo mass can also be inferred by measuring the clustering of DLAs, through their cross-correlation with the Lyman- α forest (Font-Ribera et al. 2012). Present clustering data indicates a halo mass moderately larger than anticipated by simulations (although see Bird et al. 2015).

Metal absorbers have been detected associated with DLAs (Rafelski et al. 2012). DLAs show a wide range of

metallicity, decreasing at higher redshift, with a median value which is subsolar (Jorgenson et al. 2013; Rafelski et al. 2014; Mas-Ribas et al. 2016). The DLA metallicity is expected to be correlated with the host halo mass (Pontzen et al. 2008). DLAs show little dust reddening (Khare et al. 2012) and a small subset are amongst the most metal-poor objects known (Pettini et al. 2008). Furthermore, DLAs are usually hosted by low luminosity objects, whose low star formation rate means they are not currently observable in emission (Fumagalli et al. 2015).

In this work we present new estimates for the abundance of DLAs, the average neutral hydrogen density, and the column density distribution function at $z > 2$. These population statistics are important in the wider context of galaxy formation theory. As neutral hydrogen is the raw material for star formation, measurements of neutral hydrogen can provide an independent constraint on theoretical models tuned to match the galaxy stellar mass function.

We compute statistics using the DLA catalogue presented in Garnett et al. (2016)¹, an analysis of Sloan Digital Sky Survey III Data Release 12 (SDSS–III DR12) (Eisenstein et al. 2011; Dawson et al. 2013; Alam et al. 2015) us-

* E-mail: sbird4@jhu.edu

¹ Available at http://tiny.cc/dla_catalog_gp_dr12q_lyb

ing a set of statistical techniques based on machine learning and Gaussian processes. DR12 contains the largest currently available quasar spectroscopic survey, but to achieve throughput it sacrifices spectrograph resolution and has relatively low signal-to-noise ratio in typical measurements. These properties make detecting DLAs challenging, despite the distinctive shape of the damping wing of the Voigt profile, which is a powerful discriminant from other absorbers. Nevertheless, several methods have been used to successfully detect DLAs in SDSS spectra. These have included a visual-inspection survey (Slosar et al. 2011), visually guided Voigt profile fitting (Prochaska et al. 2005; Prochaska & Wolfe 2009), a template-matching approach (Noterdaeme et al. 2009, 2012) and an unpublished machine-learning approach using Fisher discriminant analysis (Carithers 2012). These methods together have succeeded in detecting large catalogues of DLAs, but each relies on a labour-intensive visual component that limits their ability to scale to larger surveys. The methods described in Garnett et al. (2016), which use these prior catalogues as training data, are fully automated, allowing them to be used even for forthcoming much larger surveys without guidance.

The DLA catalogue of Garnett et al. (2016) is for the first time able to provide a probability of the presence of a DLA in each quasar spectrum, allowing for “soft” detections in noisy data. Furthermore, the catalogue includes a full posterior distribution for the column density and redshift of a putative DLA. This allows us to propagate uncertainties from the inference performed on each spectrum into the global population. In particular, we are often uncertain as to whether a particular low signal-to-noise spectrum contains a DLA. This will be reflected in the computed probability of that spectrum containing a DLA, allowing us to account for this uncertainty without biasing our final results. Thus we are able to include the full sample of low signal-to-noise DLAs in our measurements, substantially increasing the sample size.

Section 2 details how we estimate the statistical properties of the DLAs from the catalogue. Then in Section 3 we present our results. Section 3.1 discusses the column density distribution function (CDDF), while Section 3.2 gives the line abundance of DLAs per unit pathlength and the average universal density in neutral hydrogen. We present null checks for systematic error in Section 4; Section 4.1 shows independence of our result to signal-to-noise ratio, while Section 4.2 discusses a residual sensitivity to quasar redshift. We conclude in Section 5.

2 METHODS

DLAs are detected using the method presented in Garnett et al. (2016). This method uses Bayesian model selection, with priors derived from SDSS DR9 (Lee et al. 2013). It produces a probability that each DR12 spectrum contains a DLA as well as a full posterior probability distribution for the DLA redshift and column density in the case that it does contain a DLA. Here we present a pedagogical summary of the approach, deferring to Garnett et al. (2016) for further details. All our code is publicly available at https://github.com/rmgarnett/gp_dla_detection, and the DLA

catalogue can be found at http://tiny.cc/dla_catalog_gp_dr12q_lyb.

As DLAs are observed via absorption in quasar spectra, we need a model for the emission properties of a quasar in the absence of a DLA. This spectrum includes the quasar continuum, observational noise, and associated absorption from matter between the quasar and the observer. We model all these quantities using Gaussian processes (Rasmussen & Williams 2006). A Gaussian process is a nonparametric distribution over functions, the function space analogue of the normal distribution. To use Gaussian processes to make predictions for new data, it is necessary to impose a prior form for the mean emission and the covariance between different observational data points. Most applications of Gaussian processes use a standard, off-the-shelf form for the covariance, such as the exponential squared kernel. However, given the complex shape of quasar continua, no such standard form is appropriate. Instead, we learn a prior distribution for the shape of quasar emission without a DLA using a training sample of previously analysed quasar spectra, for which we already have DLA information. Specifically, we trained our model on the concordance DLA catalogue from SDSS DR9 (Lee et al. 2013). This concordance DLA catalogue is the “majority vote” of the DLA catalogue published as Noterdaeme et al. (2009), an unpublished DLA catalogue by Carithers (2012), and DLAs found by visual inspection (Slosar et al. 2011). Overall, it comprises a large sample, containing $\sim 50,000$ spectra. This training set enables us to estimate the prior probability and covariance of a given DLA-free quasar spectrum, which we model with a Gaussian process with mean and covariance functions we fit to the data.

We model both observational noise and absorption from matter along the path between the observer and the quasar by adding Gaussian noise contributions with appropriate variances to the emission model. The variance for the observational noise model is a function of the (observer-frame) wavelength of the spectrum, and given by the SDSS pipeline noise estimates. The variance for the forest absorption, corresponding to the mean flux from absorbing material, is learnt in a similar way to the emission model.

We may then transform the derived model for DLA-free quasar spectra into a model for the shape of an observed quasar spectrum containing a DLA. Gaussian processes offer a straightforward mechanism to incorporate simulated absorption profiles of a putative DLA. Each DLA is parametrised by its redshift and column density; given these parameters, we may evaluate the likelihood of the data under the alternative DLA model. This likelihood will be high when the simulated absorption profile explains the observed data better than the simpler DLA-free model. To compute the evidence of this model given the data, we must marginalise the a priori unknown DLA parameters under a chosen prior distribution. The DLA redshift is a priori assumed to be distributed uniformly between a point 3000 km/s blue-wards of the quasar emission redshift (to avoid a possible quasar proximity zone) and the wavelength of Lyman- β emission in the quasar rest frame (to avoid absorption from the Lyman- β forest)². The column density, N_{HI} ,

² Note that this differs slightly from the catalog of Garnett et al.

has a prior distribution given by a mixture between the column density function from the training set and, to allow for measurement uncertainty, a uniform prior on $\log N_{\text{HI}}$.

We now have two models for the observed spectrum of the quasar, one modelling the emission in the absence of a DLA, the other in the presence of a DLA. We would like to compare how well they explain a given spectrum, and thus find the posterior probability of a DLA with a particular column density or redshift along the line of sight. To do this we evaluate the Bayesian evidence for each model, assuming that the prior probability of a DLA is uniform in quasar redshift path length, and equal to the overall fraction of DLAs per unit redshift path in the training set. This integral is non-analytic, so we estimate it via numerical quadrature with 10 000 realizations of DLA parameters covering the space exhaustively. The posterior distribution over DLA parameters may be derived via normalization.

Our ability to obtain the full posterior probability distribution for the properties of a putative DLA is a strong advantage of our technique when estimating derived properties of the DLA population. We are able to use this posterior distribution to draw inferences about the properties of the overall DLA population even from quite noisy data, which may leave considerable uncertainty in the properties of any particular DLA or quasar. We will propagate this uncertainty through the analysis, ultimately allowing us to derive stronger statistical power than was available previously.

Refinements in future work might involve modifying our analysis pipeline to include data from metal lines associated with the DLA and thus better constrain the redshift and column densities of individual absorbers.

2.1 Estimating DLAs in a Column Density Bin

In this section, we explain how we turn our measured posterior distribution of DLAs, determined as explained above and in [Garnett et al. \(2016\)](#), into a measurement of the average binned column density distribution function (CDDF), the abundance of DLAs with redshift (dN/dX) and average matter density in DLAs (Ω_{DLA}). To do this we will first compute the expected number of DLAs in a given column density and redshift bin. This involves computation of summed products of probabilities from all spectra within the sample. We will show that this computation can be done directly with a few simple approximations.

The DLA pipeline provides, for each spectrum, the probability that it contains a DLA, p_{DLA}^i , and the likelihood of each sample in the (N_{HI}, z) space. The probability of a DLA in a given redshift and column density range in a particular spectrum is the sum of the likelihoods for all samples in the range, multiplied by the probability of the spectrum containing a DLA.

The total number of DLAs in the entire spectral sample in a given bin, N , is the sum of n binomial processes (the Poisson–binomial distribution), and so the probability that there are N DLAs is:

$$P(N) = \sum_{n=N} \prod_{i \in \text{DLA}} p_{\text{DLA}}^i \prod_{i \notin \text{DLA}} (1 - p_{\text{DLA}}^i). \quad (1)$$

(2016), who used the wavelength of the Lyman limit as a lower bound.

Here p_{DLA}^i is the probability that the i th spectrum contains a DLA. The sum is over all possible choices of N DLAs from n possible spectra. For each choice, there are N spectra that contain a DLA and $n - N$ which do not. The first product is over the N spectra which do contain a DLA in this draw of N DLAs from n spectra. The second product is over all other spectra, the $n - N$ which do not contain a DLA in this draw.

For the large number of spectra we consider here, it is too numerically intensive to evaluate this function directly. [Le Cam \(1960\)](#) showed that the Poisson–binomial distribution can be approximated as a Poisson distribution for small p . To take advantage of this, we split the spectra into two sets, based on whether the probability of a DLA in this bin is greater or less than a transition probability, p_{switch} . If $p < p_{\text{switch}}$, we approximate the Poisson–binomial as a Poisson distribution with mean $\sum p_{\text{DLA}}^i$. For spectra with $p > p_{\text{switch}}$, Eq. 1 is evaluated directly using a discrete Fourier transform ([Fernandez & Williams 2010](#)). The two probability distributions are then combined and the confidence intervals computed. Following [Le Cam \(1960\)](#), we use $p_{\text{switch}} = 0.25$, but we have verified that our results are unchanged with $p_{\text{switch}} = 0.5$. We also neglect (that is, assume they never contain a DLA) spectra where $p_{\text{DLA}}^i < 0.05$, as well as individual samples with $p_{\text{DLA}}^i < 10^{-4}$, as an optimisation. We have verified that lower probability spectra and samples make a negligible difference to our results.

This procedure allows us to compute the posterior probability distribution of the total number of DLAs in the sample. The confidence intervals thus reflect the uncertainty from DLA detections and redshift and column density estimation. There is also a sampling error, arising from the finite sampling of space by quasar sightlines. However, due to the large sample provided by SDSS–III, this error is very small. We have verified this by resampling our catalogue multiple times with replacement (bootstrap errors). The central value and spread of the results was extremely similar to the full confidence limits, verifying that sample variance is indeed subdominant.

We define the column density distribution, $f(N)$, such that $f(N)$ is the number of absorbers per unit column density per unit absorption distance with column density in the interval $[N, N + dN]$. Thus

$$f(N) = \frac{F(N)}{\Delta N \Delta X(z)}, \quad (2)$$

where $F(N)$ is the number of absorbers along simulated sightlines in a given column density bin. $\Delta X(z)$ is the absorption distance per sightline, defined to account for evolution in line number density with the Hubble flow:

$$X(z) = \int_0^z (1 + z')^2 \frac{H_0}{H(z')} dz'. \quad (3)$$

Here the Hubble function is $H^2(z)/H_0^2 = \Omega_{\text{M}}(1 + z)^3 + \Omega_{\Lambda}$.

We also show the incident rate of DLA systems, $\frac{dN}{dX}$, and the total HI density in DLAs, $\Omega_{\text{DLA}} \cdot \frac{dN}{dX}$ is the integral of the CDDF:

$$\frac{dN}{dX} = \int_{10^{20.3}}^{\infty} f(N, X) dN, \quad (4)$$

while Ω_{DLA} is the integral of its first moment:

$$\Omega_{\text{DLA}} = \frac{m_{\text{P}} H_0}{c \rho_{\text{c}}} \int_{10^{20.3}}^{\infty} N f(N, X) dN. \quad (5)$$

Here ρ_c is the critical density at $z = 0$ and m_p is the proton mass. Note that this definition differs by a factor of $X_H = 0.76$, the primordial hydrogen mass fraction, from the quantity quoted by [Noterdaeme et al. \(2012\)](#). The difference comes from the fact that they divide the total mass of neutral hydrogen by the hydrogen mass fraction in order to obtain the total gas mass in DLAs. However, this assumes the neutral fraction of hydrogen is unity, while in fact some of the hydrogen in DLAs may be molecular, implying that total gas mass in DLAs is greater than $\Omega_{\text{DLA}}/0.76$ by an uncertain amount. We therefore quote the total HI mass, and have adjusted the results from [Noterdaeme et al. \(2012\)](#) to match our convention. Because the CDDF falls off sharply above the DLA threshold, $\frac{dN}{dX}$ is dominated by systems with $N_{\text{HI}} \sim 5 \times 10^{20} \text{ cm}^{-2}$, while Ω_{DLA} is strongly affected by higher column densities with $N_{\text{HI}} \sim 10^{21} \text{ cm}^{-2}$.

Note that the [Noterdaeme et al. \(2012\)](#) measurements of $\frac{dN}{dX}$ do not include error bars, as the error was dominated by the uncertainty in detections of weak DLAs (P. Noterdaeme, private communication). An advantage of our method is that we can quantify this uncertainty, and thus we do include error bars on $\frac{dN}{dX}$.

We omit from our catalogue the regions of spectra potentially affected by Lyman- β absorption, that is, those with a rest wavelength of $\lambda < 1026.72 \text{ \AA}$. As our simple model for the Lyman- α forest absorption did not include Lyman- β , the Gaussian process code misidentifies some of this absorption (especially in the most noisy spectra) as potential DLAs. The redshift distribution of SDSS quasars peaks at relatively low redshift, meaning that the total fraction of path length affected is relatively small. Thus we simply regenerate the catalog while excising the affected part of the spectrum. An alternative method would be to explicitly model absorption from Lyman- β , something we may attempt in future work.

We assume that each spectrum contains at maximum one DLA. By neglecting spectra containing two strong DLAs, this may marginally decrease the dN/dX over that observed. The most detectable DLA in a spectrum is likely to occur at low redshift, so the objects missed will be more common at high redshift. However, DLAs are relatively rare (and the spectrum we examine limited to that between 1216 and 1027 \AA) so this is unlikely to be a significant effect. Note this only matters for spectra containing two objects that are overwhelmingly likely to be DLAs, ie, where the expected number of DLAs in the spectrum is greater than one. A spectrum containing two objects with $p_{\text{DLA}} \sim 0.1$, for example, will be counted correctly as $p_{\text{DLA}} \sim 0.2$.

When computing Ω_{DLA} , we must compute not only the number of DLAs in a redshift bin, but also their total column density. For computational reasons, we estimate this by constructing the column density function in each redshift bin of interest and then using Eq. 5 to obtain Ω_{DLA} by integration. For each redshift bin we compute the full posterior PDF for 30 column density bins, using the method described above. These (discrete) PDFs are then multiplied together to estimate the posterior distribution for the total matter density at each redshift.

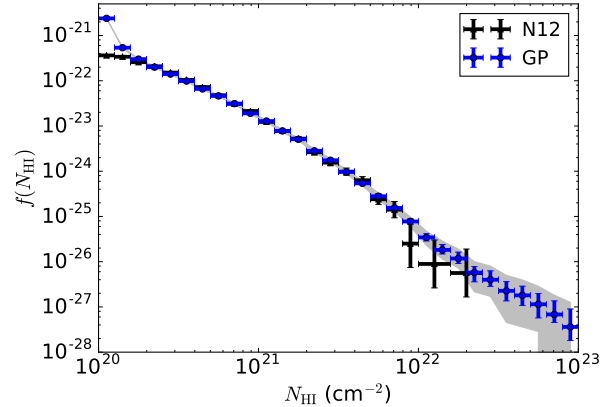


Figure 1. The CDDF, integrated over all $z < 5$ spectral path, derived from our DR12 DLA catalogue (blue, labelled ‘GP’) and compared to the CDDF derived from the DLA catalogue of [Noterdaeme et al. \(2012\)](#) (N12; black). 68% confidence limits are shown as error bars, while 95% confidence limits are shown as a grey filled band. Note that at low column densities our results completely overlap those of N12, and the large sample means that errors are extremely small.

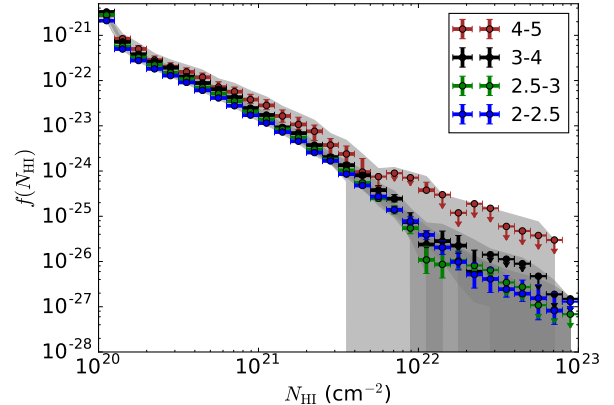


Figure 2. The CDDF derived from DLAs in a variety of redshift ranges. Labels show the redshift range of the absorbers used. 68% confidence limits are shown as error bars, while 95% confidence limits are shown as a grey filled band. In column density and redshift ranges where there was no detection at 68% confidence, a down-pointing arrow is shown indicating the 68% confidence upper limit.

3 RESULTS

3.1 Column Density Distribution Function

Figure 1 shows the CDDF from our analysis of DR12, compared to the DR9 DLA catalogue from [Noterdaeme et al. \(2012\)](#). We combine all spectral path where an absorber would be at $z < 5$ to obtain a path-weighted average over redshift. In practice for SDSS DR12 the total path is dominated by $z < 2.5$. Error bars denote 68% confidence limits, while the grey band shows the region enclosed by the 95% confidence limits. Table A1 contains our CDDF in tabulated form.

The lowest two column density bins are somewhat high due to an artefact of our analysis method. Our column density prior for each spectrum is cut off at $N_{\text{HI}} < 10^{20}$, but the

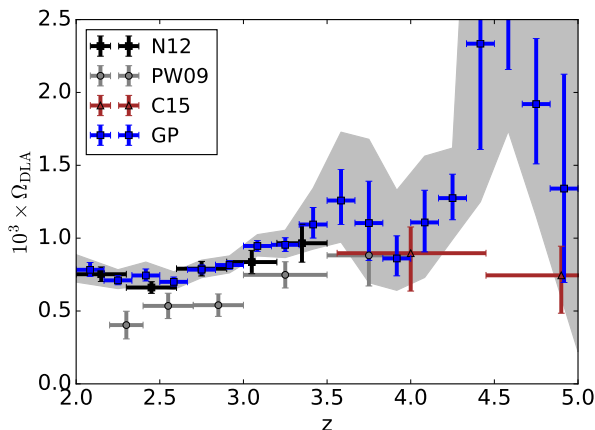


Figure 3. The HI density in DLAs, Ω_{DLA} , from our DR12 DLA catalogue as a function of redshift (blue, labelled ‘GP’), compared to the results of [Noterdaeme et al. \(2012\)](#) (N12; black), [Prochaska & Wolfe \(2009\)](#) (PW09; grey) and [Crighton et al. \(2015\)](#) (C15; red). See text for an explanation of the mild tension between measurements.

lower limit is purely a matter of the conventional definition of a DLA. Our detection method is capable of distinguishing somewhat weaker absorbers from noise, and correctly assigns them a maximum posterior column density corresponding to the lowest value allowed by the prior. Note this will not affect our results for dN/dX or Ω_{DLA} , because by definition these only include DLAs, i.e., samples with $N_{\text{HI}} > 10^{20.3}$. We may perform further analysis of these weak absorbers in future work, effectively extending our prior to lower column densities.

Our results agree well with those of [Noterdaeme et al. \(2012\)](#). Our error bars are substantially smaller, emphasising the increased statistical size of our sample. This is due not just to the larger size of DR12, but also to the ability of our method to quantify the uncertainty in a DLA measurement. This allows us to robustly include noisier spectra into our estimate, greatly increasing the size of the available sample. Interestingly the data shows no indication of a molecular hydrogen cutoff at high column density ([Altay et al. 2011](#)). We detect the presence of DLAs at 95% confidence for $N_{\text{HI}} < 6 \times 10^{22} \text{ cm}^{-2}$, and at 68% confidence for $N_{\text{HI}} < 10^{23} \text{ cm}^{-2}$.

Figure 2 shows the CDDF in a variety of redshift bins. Figure 2 indicated by downward pointing arrows show the 68% upper confidence limit when the data is consistent with zero at 68% confidence. We detect minimal evolution in the shape of the CDDF with redshift for $z < 4$. The CDDF amplitude gently increases with redshift, as we shall discuss in more detail by looking at dN/dX and Ω_{DLA} in Section 3.2. Interestingly, there is some evidence of the slope of the CDDF becoming shallower at $z > 4$, although we do not yet reliably detect DLAs with column density $N_{\text{HI}} > 3 \times 10^{21} \text{ cm}^{-2}$. It will be interesting to see if larger data sets confirm a shallower slope, or point to a cutoff.

3.2 Redshift Evolution of DLAs

Figure 3 shows Ω_{DLA} , the HI density in DLAs in units of the cosmic density (see also Table A6). Our results are consistent with those of [Noterdaeme et al. \(2012\)](#). As dis-

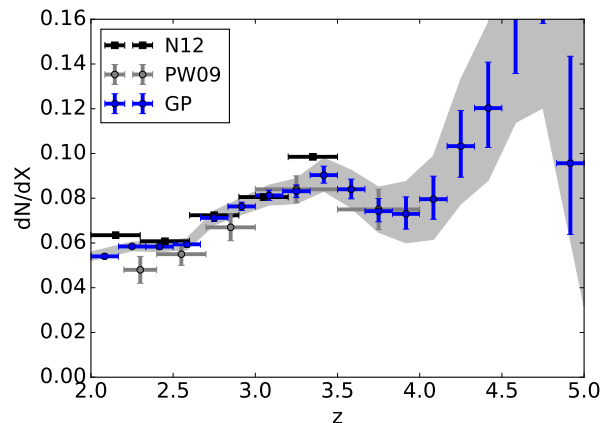


Figure 4. The line density of DLAs as a function of redshift from our DR12 DLA catalogue (blue, ‘GP’), compared to the results of [Noterdaeme et al. \(2012\)](#) (N12; black) and [Prochaska & Wolfe \(2009\)](#) (PW09; grey). Note that statistical error was not computed in [Noterdaeme et al. \(2012\)](#).

cussed in [Noterdaeme et al. \(2012\)](#) and [Sánchez-Ramírez et al. \(2016\)](#), the discrepancy between [Noterdaeme et al. \(2012\)](#) and [Prochaska & Wolfe \(2009\)](#) is driven by the relatively small size of the SDSS DR5 dataset used in [Prochaska & Wolfe \(2009\)](#). In DR5 there were relatively large statistical errors for $N_{\text{HI}} > 3 \times 10^{21} \text{ cm}^{-2}$, and a statistical fluctuation caused these bins to be lower than revealed by later datasets.

We also show, for the first time from SDSS, measurements of Ω_{DLA} at $z > 4$. As in previous Figures, we include data only for $z < 5$, as at higher redshifts our 95% confidence limits are consistent with zero DLAs. Due to the relatively small sample (and noisy spectra) at $z > 4$, we do not convincingly detect DLAs with $N_{\text{HI}} > 3 \times 10^{21} \text{ cm}^{-2}$ at this redshift (see Figure 2). For column densities $2 \times 10^{20} < N_{\text{HI}} < 10^{21}$, our CDDF is roughly a power law with slope -1.7 , and for $N_{\text{HI}} > 10^{21}$ it steepens and has a slope roughly -2.5 . Thus we a priori expect column densities with $N_{\text{HI}} > 3 \times 10^{21} \text{ cm}^{-2}$ to contribute around 20% of Ω_{DLA} . This expectation is incorporated in the prior CDDF applied to spectra where it is not possible to precisely estimate the DLA column density, and implies that a future, larger, sample, should decrease our statistical uncertainty, but not increase the measured value of Ω_{DLA} .

Figure 3 compares our measurements to those of [Crighton et al. \(2015\)](#), who used a smaller survey of high signal to noise spectra. While our data prefers a slightly higher value of Ω_{DLA} than theirs, this discrepancy is smaller than it appears. Like us, [Crighton et al. \(2015\)](#) securely detect DLAs to $N_{\text{HI}} > 3 \times 10^{21} \text{ cm}^{-2}$, but, when computing Ω_{DLA} , they do not include a contribution from DLAs with column densities larger than they detect. If we approximate their method by integrating Eq. 5 only to the largest column density detected in our dataset, our measurement of Ω_{DLA} reduces by about 1.5σ at $z > 3.5$, completely reconciling our results.

Figure 4 shows the line density of DLAs (see Table A6). Here our results are consistent with those of [Prochaska & Wolfe \(2009\)](#) and [Noterdaeme et al. \(2012\)](#) where these two agree. Where they disagree we are between the two at low redshift and consistent with [Prochaska & Wolfe \(2009\)](#) at

high redshift. dN/dX is sensitive to the weakest DLAs, making some earlier methods prone to false positives, especially at high redshifts where there is strong absorption from the Lyman- α forest. Indeed, [Noterdaeme et al. \(2012\)](#) perform a substantial correction to their highest redshift bin using mocks, and our results suggest this correction was moderately too small. Our method is robust to this problem; in the presence of strong absorption from the forest it simply reduces the strength of the detection. We see that the template fitting of [Noterdaeme et al. \(2012\)](#) also overestimates dN/dX in the lowest redshift bin. This may be due to the difficulty of securely detecting DLAs using templates when the spectrum is too short to include the full damping wing.

Both Ω_{DLA} and dN/dX exhibit a similar pattern of evolution with redshift. From $z = 2 - 2.75$ they evolve minimally, if at all. Between $z = 2.75$ and $z = 3.5$ they both increase gently to higher redshifts. This increase is now detected at fairly high significance, and is a little curious as it does not appear to be reproduced in cosmological simulations (e.g. [Bird et al. 2014](#); [Rahmati et al. 2015](#)).

A constant Ω_{DLA} at $z > 3.5$ lies close to our 95% confidence limits, suggesting that what evolution appears may be statistical. However, dN/dX shows a noticeable dip between $z = 3.5$ and $z = 4.5$, and an increase for $z > 4.5$. We extracted dN/dX from the simulations used in [Bird et al. \(2014\)](#), and found that it was roughly constant at these redshifts. One possibility is that at these high redshifts, where strong Lyman- α forest absorption makes estimating the neutral hydrogen column density more difficult, we are mistaking lower column density systems for DLAs. Another is that this is simply a rare statistical fluctuation.

One possible explanation is selection effects from SDSS colour selection. The quasar rest-frame Lyman limit at 912 Å moves into the observable region of the spectrum at $z = 3.7$. [Prochaska et al. \(2009\)](#) identified a systematic in the SDSS colour selection which causes quasar sightlines containing Lyman Limit Systems (LLS) to be preferentially selected for spectroscopic followup. [Worseck & Prochaska \(2011\)](#) and [Fumagalli et al. \(2013\)](#) showed that, due to the width of the u -band filter in SDSS, LLS are over-sampled for $z = 2.5 - 3.6$ for all quasars in the redshift range $z = 3.0 - 3.6$. In order to determine whether this systematic was affecting our results, we considered a modified catalogue with all affected spectral path excised. We found that the feature at $z = 3.5 - 4.5$ is also present in $z > 3.6$ quasars, which do not suffer from this bias. Thus the feature persists in a clean sample, albeit at reduced significance. LLS are more common than DLAs and so the correlation between them is weak enough that a biased selection of LLS apparently does not bias the abundance of DLAs. We note however that other, as yet undiscovered, systematics due to colour selection remain possible.

It is interesting that, although they note that their measurement of the line density has significant systematic uncertainty, [Crighton et al. \(2015\)](#) also observe an increase in the line density of DLAs between $z = 4$ and $z = 4.9$, to $dN/dX \sim 0.1$, in agreement with our results. It will be interesting to see whether this trend remains in future datasets.

4 CHECKS FOR SYSTEMATICS

4.1 Effect of SNR

Figure 5 shows the DLA abundance from subsets of the DR12 quasar catalogue with signal to noise (SNR) > 2 and > 4 . This demonstrates that our results are insensitive to the SNRs of the spectra. We define the SNR as the median of the ratio between the flux and the pixel noise within spectrum redwards of the rest-frame Lyman- α emission. We do not use the Lyman- α part of the spectrum, to avoid inducing correlations between SNR and the presence of a DLA. Since our definition uses a part of the spectrum in which there is strong Lyman- α emission and little absorption, it produces a higher SNR than the usual definition using the noise in the forest. Under our definition, 70% of spectra have SNR > 8 , 92.5 % SNR > 4 , and 98 % SNR > 2 . The effect of restricting to high SNR spectra is extremely small, less than one half the 68% confidence limits of the measurement, verifying that our results are not biased by noise.

We have also verified explicitly that our results are not sensitive to the noise in a given pixel. To do this, we omitted from our catalogue spectral pixels where the signal to noise in the pixel was less than 2; the derived values for Ω_{DLA} and dN/dX changed by less than the 68% confidence limits.

4.2 Quasar Redshift

Due to the large distance separating absorbers from the background quasar, the properties of the absorber should be largely uncorrelated with the properties of the background quasar illuminating it. However, many common sources of systematic error will occur in the rest frame of the quasar. Figure 6 shows both dN/dX and Ω_{DLA} for our dataset binned as a function of quasar redshift. In the absence of systematics, we would expect both quantities to be the same irrespective of the redshift of the quasar, and thus the two lines should overlap. This is an extremely powerful test; for example, if the Lyman- β forest part of the spectrum is not removed, the effect of Lyman- β absorption is highly visible.

As shown by Figure 6, there is some residual correlation between DLA properties and quasar redshift, in particular in the line density of the DLAs at $z < 3$. The majority of this correlation comes from the area nearest the quasar. We found that removing the spectrum with $z - z_{\text{QSO}} < 0.1$ was sufficient to remove all correlation between the quasar redshift and Ω_{DLA} . Some correlation still remained between dN/dX and z_{QSO} at $z < 2.5$. At $z = 3$, $\Delta z = 0.1$ corresponds to a physical distance of ~ 1 Gpc, much larger than the expected mean free path to ionizing photons, ~ 100 Mpc (e.g. [Haardt & Madau 2012](#)). This signal is thus unlikely to be an extended proximity zone around quasars. One possibility is some systematic in the data; perhaps a small subset of quasars in this dataset has sufficient emission near the redshift of the quasar that our Gaussian process code fails to detect the DLA. If such objects were sufficiently rare, they might not appear in the training set. However, it seems unlikely that an effect of this magnitude could be missed from the training set. Another possibility is some interaction between DLAs and the SDSS colour selection algorithm which mildly biases spectroscopic followup towards quasars which

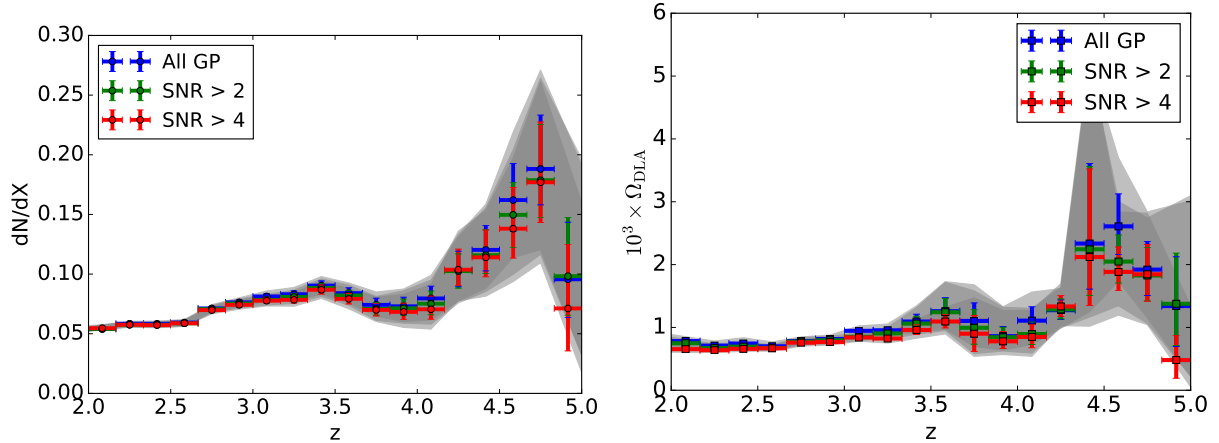


Figure 5. The line density and Ω_{DLA} in DLAs as a function of redshift for a sample with different minimal signal to noise ratios. The green points show $\text{SNR} > 2$, which exclude 2% of the noisiest spectra. The red points show $\text{SNR} > 4$, which exclude 7.5% of the noisiest spectra. Grey bands show the 95% confidence intervals, which overlap very closely for all samples.

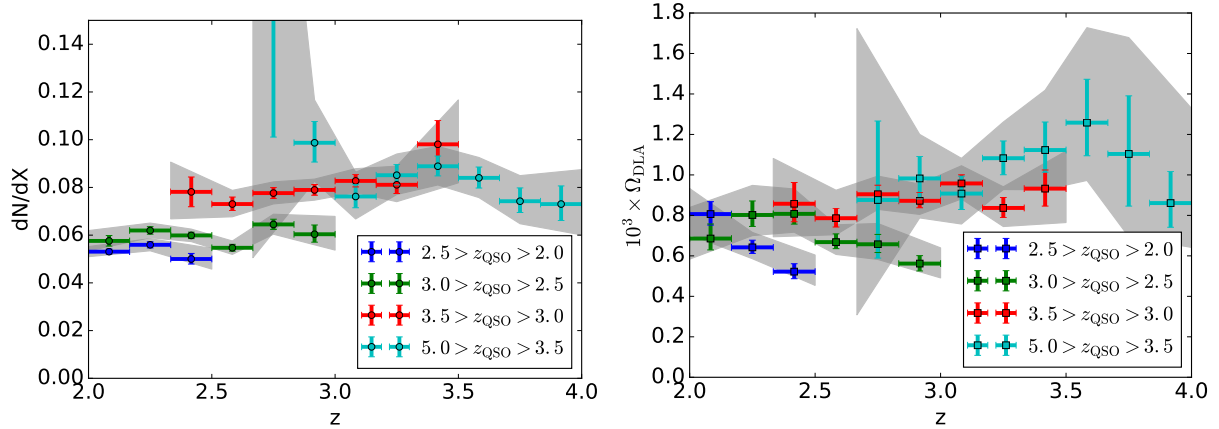


Figure 6. Redshift evolution of the line density in DLAs (Left) and matter density in DLAs, Ω_{DLA} (Right), as a function of the redshift of the quasar (z_{QSO}). Correlation between the properties of the background quasar and those of the absorber may indicate systematic error.

do not contain a DLA near the redshift of the quasar³, or a systematic bias in quasar redshift.

There are also potential explanations which do not involve observational systematics. For example, quasars are clustered, inhabiting overdense regions. Thus we may be detecting some correlation with the gas in these overdense regions. This could be either a reduction in gas corresponding to the more advanced evolution of structure growth in the overdense region, or overlapping proximity regions of nearby faint or obscured quasars.

We have investigated removing the region near the quasar redshift, and found that our overall results change by about the 68% confidence limits. As it is unclear whether the explanation for this discrepancy is observational or physical, we choose to estimate DLA properties using all available data and caution the reader that there may be unmodelled

systematics of the approximate magnitude of the 68% confidence limits.

5 CONCLUSION

We have presented a new measurement of the abundance of neutral hydrogen from $z = 2$ to $z = 5$, using a catalogue built from SDSS DR12 spectra (Garnett et al. 2016). Our results are in good agreement with previous measurements by Noterdaeme et al. (2012), Prochaska & Wolfe (2009) and Crighton et al. (2015). The robust statistical properties of our method allow us to include partial detections of DLAs in noisy spectra, which allows us to use even the noisiest spectra and thus substantially increase the size of the sample. We note, however, that there is a residual dependence on quasar redshift for $z < 2.5$, which may indicate un-modelled systematic errors within our data, of similar magnitude to the statistical uncertainty. Our larger sample size allows us to substantially reduce statistical error and extend our measurement to $z > 4$ for the first time using SDSS spectra. We

³ Note that the colour selection bias towards LLS described in Worseck & Prochaska (2011) does not affect absorbers at $z < 2.5$ and so cannot provide a full explanation.

detect no evolution in the shape of the CDDF for $2 < z < 4$, while the abundance of both DLAs and neutral hydrogen increases moderately over the same redshift range.

ACKNOWLEDGEMENTS

SB was supported by a McWilliams Fellowship from Carnegie Mellon University and by NASA through Einstein Postdoctoral Fellowship Award Number PF5-160133. SH was supported by NASA and NSF for this work.

APPENDIX A: TABLES OF RESULTS

REFERENCES

- Alam S., et al., 2015, [ApJS](#), **219**, 12
- Altay G., Theuns T., Schaye J., Crighton N. H. M., Dalla Vecchia C., 2011, [ApJ](#), **737**, L37
- Bird S., Vogelsberger M., Haehnelt M., Sijacki D., Genel S., Torrey P., Springel V., Hernquist L., 2014, [MNRAS](#), **445**, 2313
- Bird S., Haehnelt M., Neeleman M., Genel S., Vogelsberger M., Hernquist L., 2015, [MNRAS](#), **447**, 1834
- Carithers W., 2012, DLA Concordance Catalog, Published internally to SDSS
- Cen R., 2012, [ApJ](#), **748**, 121
- Crighton N. H. M., et al., 2015, [MNRAS](#), **452**, 217
- Dawson K. S., et al., 2013, [AJ](#), **145**, 10
- Eisenstein D. J., et al., 2011, [AJ](#), **142**, 72
- Fernandez M., Williams S., 2010, [Aerospace and Electronic Systems, IEEE Transactions on](#), **46**, 803
- Font-Ribera A., et al., 2012, [JCAP](#), **11**, 059
- Fumagalli M., O’Meara J. M., Prochaska J. X., Worseck G., 2013, [ApJ](#), **775**, 78
- Fumagalli M., O’Meara J. M., Prochaska J. X., Rafelski M., Kanekar N., 2015, [MNRAS](#), **446**, 3178
- Gardner J. P., Katz N., Weinberg D. H., Hernquist L., 1997, [ApJ](#), **486**, 42
- Garnett R., Ho S., Bird S., Schneider J., 2016, preprint, ([arXiv:1605.04460](#))
- Haardt F., Madau P., 2012, [ApJ](#), **746**, 125
- Haehnelt M. G., Steinmetz M., Rauch M., 1998, [ApJ](#), **495**, 647
- Jorgenson R. A., Murphy M. T., Thompson R., 2013, [MNRAS](#), **435**, 482
- Khare P., vanden Berk D., York D. G., Lundgren B., Kulkarni V. P., 2012, [MNRAS](#), **419**, 1028
- Le Cam L., 1960, *Pacific J. Math.*, **10**, 1181
- Lee K.-G., et al., 2013, [AJ](#), **145**, 69
- Mas-Ribas L., et al., 2016, preprint, ([arXiv:1610.02711](#))
- Noterdaeme P., Petitjean P., Ledoux C., Srianand R., 2009, [A&A](#), **505**, 1087
- Noterdaeme P., et al., 2012, [A&A](#), **547**, L1
- Pettini M., Zych B. J., Steidel C. C., Chaffee F. H., 2008, [MNRAS](#), **385**, 2011
- Pontzen A., et al., 2008, [MNRAS](#), **390**, 1349
- Prochaska J. X., Wolfe A. M., 1997, [ApJ](#), **487**, 73
- Prochaska J. X., Wolfe A. M., 2009, [ApJ](#), **696**, 1543
- Prochaska J. X., Herbert-Fort S., Wolfe A. M., 2005, [ApJ](#), **635**, 123
- Prochaska J. X., Worseck G., O’Meara J. M., 2009, [ApJ](#), **705**, L113
- Rafelski M., Wolfe A. M., Prochaska J. X., Neeleman M., Mendez A. J., 2012, [ApJ](#), **755**, 89
- Rafelski M., Neeleman M., Fumagalli M., Wolfe A. M., Prochaska J. X., 2014, [ApJ](#), **782**, L29
- Rahmati A., Schaye J., Bower R. G., Crain R. A., Furlong M., Schaller M., Theuns T., 2015, [MNRAS](#), **452**, 2034
- Rasmussen C. E., Williams C. K. I., 2006, *Gaussian Processes for Machine Learning*. MIT Press
- Sánchez-Ramírez R., et al., 2016, [MNRAS](#), **456**, 4488
- Slosar A., et al., 2011, [JCAP](#), **9**, 1
- Wolfe A. M., Turnshek D. A., Smith H. E., Cohen R. D., 1986, [ApJS](#), **61**, 249
- Worseck G., Prochaska J. X., 2011, [ApJ](#), **728**, 23
- Zafar T., Péroux C., Popping A., Milliard B., Deharveng J.-M., Frank S., 2013, [A&A](#), **556**, A141

$\log_{10} N_{\text{HI}}$	$f(N_{\text{HI}}) (10^{-21})$	68% limits (10^{-21})	95% limits (10^{-21})
20.0 – 20.1	2.392	2.377 – 2.408	2.362 – 2.423
20.1 – 20.2	0.538	0.531 – 0.546	0.523 – 0.554
20.2 – 20.3	0.300	0.295 – 0.305	0.290 – 0.310
20.3 – 20.4	0.200	0.197 – 0.204	0.193 – 0.208
20.4 – 20.5	0.141	0.138 – 0.143	0.135 – 0.146
20.5 – 20.6	9.84×10^{-2}	$[9.64 - 10.05] \times 10^{-2}$	$[9.45 - 10.24] \times 10^{-2}$
20.6 – 20.7	6.60×10^{-2}	$[6.46 - 6.75] \times 10^{-2}$	$[6.32 - 6.89] \times 10^{-2}$
20.7 – 20.8	4.62×10^{-2}	$[4.52 - 4.73] \times 10^{-2}$	$[4.41 - 4.84] \times 10^{-2}$
20.8 – 20.9	3.13×10^{-2}	$[3.05 - 3.21] \times 10^{-2}$	$[2.98 - 3.28] \times 10^{-2}$
20.9 – 21.0	1.90×10^{-2}	$[1.85 - 1.96] \times 10^{-2}$	$[1.80 - 2.01] \times 10^{-2}$
21.0 – 21.1	1.29×10^{-2}	$[1.25 - 1.33] \times 10^{-2}$	$[1.21 - 1.36] \times 10^{-2}$
21.1 – 21.2	7.79×10^{-3}	$[7.52 - 8.06] \times 10^{-3}$	$[7.27 - 8.32] \times 10^{-3}$
21.2 – 21.3	5.08×10^{-3}	$[4.90 - 5.27] \times 10^{-3}$	$[4.72 - 5.45] \times 10^{-3}$
21.3 – 21.4	2.83×10^{-3}	$[2.71 - 2.96] \times 10^{-3}$	$[2.60 - 3.08] \times 10^{-3}$
21.4 – 21.5	1.75×10^{-3}	$[1.67 - 1.83] \times 10^{-3}$	$[1.59 - 1.91] \times 10^{-3}$
21.5 – 21.6	9.70×10^{-4}	$[9.20 - 10.29] \times 10^{-4}$	$[8.70 - 10.78] \times 10^{-4}$
21.6 – 21.7	5.40×10^{-4}	$[5.08 - 5.79] \times 10^{-4}$	$[4.75 - 6.15] \times 10^{-4}$
21.7 – 21.8	2.80×10^{-4}	$[2.57 - 3.06] \times 10^{-4}$	$[2.34 - 3.29] \times 10^{-4}$
21.8 – 21.9	1.54×10^{-4}	$[1.39 - 1.70] \times 10^{-4}$	$[1.27 - 1.86] \times 10^{-4}$
21.9 – 22.0	7.76×10^{-5}	$[6.67 - 8.84] \times 10^{-5}$	$[5.95 - 9.74] \times 10^{-5}$
22.0 – 22.1	3.44×10^{-5}	$[2.87 - 4.16] \times 10^{-5}$	$[2.29 - 4.73] \times 10^{-5}$
22.1 – 22.2	1.82×10^{-5}	$[1.48 - 2.39] \times 10^{-5}$	$[1.14 - 2.85] \times 10^{-5}$
22.2 – 22.3	1.18×10^{-5}	$[9.04 - 16.27] \times 10^{-6}$	$[7.23 - 18.99] \times 10^{-6}$
22.3 – 22.4	5.75×10^{-6}	$[3.59 - 7.90] \times 10^{-6}$	$[2.15 - 10.05] \times 10^{-6}$
22.4 – 22.5	3.99×10^{-6}	$[2.85 - 6.27] \times 10^{-6}$	$[1.71 - 7.99] \times 10^{-6}$
22.5 – 22.6	2.27×10^{-6}	$[1.36 - 3.63] \times 10^{-6}$	$[4.53 - 49.84] \times 10^{-7}$
22.6 – 22.7	1.80×10^{-6}	$[1.08 - 2.88] \times 10^{-6}$	$[3.60 - 39.59] \times 10^{-7}$
22.7 – 22.8	1.14×10^{-6}	$[5.72 - 20.01] \times 10^{-7}$	$[2.86 - 28.59] \times 10^{-7}$
22.8 – 22.9	6.81×10^{-7}	$[4.54 - 13.63] \times 10^{-7}$	$0 - 1.82 \times 10^{-6}$
22.9 – 23.0	3.61×10^{-7}	$[1.80 - 9.02] \times 10^{-7}$	$0 - 1.26 \times 10^{-6}$

Table A1. Average column density distribution function for all DLAs with $2 < z < 5$. See also Figure 1.

$\log_{10} N_{\text{HI}}$	$f(N_{\text{HI}}) (10^{-21})$	68% limits (10^{-21})	95% limits (10^{-21})
20.0 – 20.1	2.112	2.091 – 2.132	2.072 – 2.152
20.1 – 20.2	0.498	0.488 – 0.508	0.478 – 0.518
20.2 – 20.3	0.277	0.271 – 0.284	0.264 – 0.291
20.3 – 20.4	0.181	0.177 – 0.186	0.172 – 0.191
20.4 – 20.5	0.129	0.125 – 0.132	0.122 – 0.136
20.5 – 20.6	9.10×10^{-2}	$[8.84 - 9.38] \times 10^{-2}$	$[8.59 - 9.63] \times 10^{-2}$
20.6 – 20.7	6.07×10^{-2}	$[5.88 - 6.27] \times 10^{-2}$	$[5.70 - 6.46] \times 10^{-2}$
20.7 – 20.8	4.12×10^{-2}	$[3.98 - 4.27] \times 10^{-2}$	$[3.85 - 4.41] \times 10^{-2}$
20.8 – 20.9	2.77×10^{-2}	$[2.67 - 2.88] \times 10^{-2}$	$[2.58 - 2.98] \times 10^{-2}$
20.9 – 21.0	1.72×10^{-2}	$[1.65 - 1.80] \times 10^{-2}$	$[1.58 - 1.86] \times 10^{-2}$
21.0 – 21.1	1.15×10^{-2}	$[1.10 - 1.20] \times 10^{-2}$	$[1.05 - 1.25] \times 10^{-2}$
21.1 – 21.2	7.27×10^{-3}	$[6.93 - 7.64] \times 10^{-3}$	$[6.58 - 8.01] \times 10^{-3}$
21.2 – 21.3	4.56×10^{-3}	$[4.31 - 4.82] \times 10^{-3}$	$[4.07 - 5.06] \times 10^{-3}$
21.3 – 21.4	2.56×10^{-3}	$[2.39 - 2.73] \times 10^{-3}$	$[2.24 - 2.90] \times 10^{-3}$
21.4 – 21.5	1.67×10^{-3}	$[1.56 - 1.80] \times 10^{-3}$	$[1.46 - 1.91] \times 10^{-3}$
21.5 – 21.6	8.56×10^{-4}	$[7.91 - 9.38] \times 10^{-4}$	$[7.26 - 10.11] \times 10^{-4}$
21.6 – 21.7	4.79×10^{-4}	$[4.34 - 5.31] \times 10^{-4}$	$[3.89 - 5.77] \times 10^{-4}$
21.7 – 21.8	2.73×10^{-4}	$[2.42 - 3.09] \times 10^{-4}$	$[2.16 - 3.40] \times 10^{-4}$
21.8 – 21.9	1.39×10^{-4}	$[1.23 - 1.64] \times 10^{-4}$	$[1.02 - 1.88] \times 10^{-4}$
21.9 – 22.0	7.79×10^{-5}	$[6.49 - 9.74] \times 10^{-5}$	$[5.52 - 11.04] \times 10^{-5}$
22.0 – 22.1	3.87×10^{-5}	$[3.10 - 4.90] \times 10^{-5}$	$[2.32 - 5.93] \times 10^{-5}$
22.1 – 22.2	2.05×10^{-5}	$[1.43 - 2.87] \times 10^{-5}$	$[1.02 - 3.48] \times 10^{-5}$
22.2 – 22.3	9.77×10^{-6}	$[6.51 - 16.28] \times 10^{-6}$	$[3.26 - 19.53] \times 10^{-6}$
22.3 – 22.4	5.17×10^{-6}	$[2.59 - 9.05] \times 10^{-6}$	$[1.29 - 11.63] \times 10^{-6}$
22.4 – 22.5	4.11×10^{-6}	$[2.05 - 7.19] \times 10^{-6}$	$[1.03 - 9.24] \times 10^{-6}$
22.5 – 22.6	2.45×10^{-6}	$[1.63 - 4.89] \times 10^{-6}$	$0 - 6.53 \times 10^{-6}$
22.6 – 22.7	1.94×10^{-6}	$[1.30 - 3.89] \times 10^{-6}$	$0 - 5.18 \times 10^{-6}$
22.7 – 22.8	1.54×10^{-6}	$[5.15 - 30.88] \times 10^{-7}$	$0 - 4.12 \times 10^{-6}$
22.8 – 22.9	8.18×10^{-7}	$[4.09 - 20.44] \times 10^{-7}$	$0 - 2.86 \times 10^{-6}$
22.9 – 23.0	3.25×10^{-7}	$0 - 1.30 \times 10^{-6}$	$0 - 1.62 \times 10^{-6}$

Table A2. Column density distribution function for all DLAs with $2 < z < 2.5$. See also Figure 2.

$\log_{10} N_{\text{HI}}$	$f(N_{\text{HI}}) (10^{-21})$	68% limits (10^{-21})	95% limits (10^{-21})
20.0 – 20.1	2.824	2.793 – 2.856	2.763 – 2.887
20.1 – 20.2	0.570	0.554 – 0.586	0.540 – 0.600
20.2 – 20.3	0.324	0.314 – 0.335	0.305 – 0.345
20.3 – 20.4	0.222	0.215 – 0.230	0.208 – 0.237
20.4 – 20.5	0.149	0.144 – 0.155	0.139 – 0.160
20.5 – 20.6	0.108	0.104 – 0.112	$[9.99 - 11.59] \times 10^{-2}$
20.6 – 20.7	7.01×10^{-2}	$[6.72 - 7.31] \times 10^{-2}$	$[6.45 - 7.59] \times 10^{-2}$
20.7 – 20.8	5.12×10^{-2}	$[4.90 - 5.34] \times 10^{-2}$	$[4.70 - 5.55] \times 10^{-2}$
20.8 – 20.9	3.51×10^{-2}	$[3.35 - 3.67] \times 10^{-2}$	$[3.21 - 3.82] \times 10^{-2}$
20.9 – 21.0	2.08×10^{-2}	$[1.98 - 2.19] \times 10^{-2}$	$[1.88 - 2.30] \times 10^{-2}$
21.0 – 21.1	1.39×10^{-2}	$[1.31 - 1.46] \times 10^{-2}$	$[1.24 - 1.54] \times 10^{-2}$
21.1 – 21.2	8.38×10^{-3}	$[7.86 - 8.94] \times 10^{-3}$	$[7.38 - 9.46] \times 10^{-3}$
21.2 – 21.3	5.45×10^{-3}	$[5.11 - 5.83] \times 10^{-3}$	$[4.80 - 6.21] \times 10^{-3}$
21.3 – 21.4	3.00×10^{-3}	$[2.78 - 3.27] \times 10^{-3}$	$[2.56 - 3.49] \times 10^{-3}$
21.4 – 21.5	1.82×10^{-3}	$[1.67 - 1.99] \times 10^{-3}$	$[1.54 - 2.14] \times 10^{-3}$
21.5 – 21.6	1.03×10^{-3}	$[9.28 - 11.35] \times 10^{-4}$	$[8.42 - 12.38] \times 10^{-4}$
21.6 – 21.7	5.46×10^{-4}	$[4.78 - 6.14] \times 10^{-4}$	$[4.23 - 6.83] \times 10^{-4}$
21.7 – 21.8	2.49×10^{-4}	$[2.17 - 3.04] \times 10^{-4}$	$[1.74 - 3.47] \times 10^{-4}$
21.8 – 21.9	1.38×10^{-4}	$[1.12 - 1.72] \times 10^{-4}$	$[9.48 - 19.82] \times 10^{-5}$
21.9 – 22.0	5.48×10^{-5}	$[4.11 - 7.53] \times 10^{-5}$	$[3.42 - 9.58] \times 10^{-5}$
22.0 – 22.1	1.09×10^{-5}	$[5.44 - 21.75] \times 10^{-6}$	$[5.44 - 32.62] \times 10^{-6}$
22.1 – 22.2	8.64×10^{-6}	$[4.32 - 21.59] \times 10^{-6}$	$0 - 2.59 \times 10^{-5}$
22.2 – 22.3	1.03×10^{-5}	$[6.86 - 17.15] \times 10^{-6}$	$[3.43 - 20.58] \times 10^{-6}$
22.3 – 22.4	2.72×10^{-6}	$0 - 8.17 \times 10^{-6}$	$0 - 1.36 \times 10^{-5}$
22.4 – 22.5	2.16×10^{-6}	$0 - 6.49 \times 10^{-6}$	$0 - 8.66 \times 10^{-6}$
22.5 – 22.6	0	$0 - 3.44 \times 10^{-6}$	$0 - 5.16 \times 10^{-6}$
22.6 – 22.7	0	$0 - 2.73 \times 10^{-6}$	$0 - 2.73 \times 10^{-6}$
22.7 – 22.8	0	$0 - 1.08 \times 10^{-6}$	$0 - 2.17 \times 10^{-6}$
22.8 – 22.9	0	$0 - 8.62 \times 10^{-7}$	$0 - 1.72 \times 10^{-6}$
22.9 – 23.0	0	$0 - 6.84 \times 10^{-7}$	$0 - 1.37 \times 10^{-6}$

Table A3. Column density distribution function for all DLAs with $2.5 < z < 3$. See also Figure 2.

$\log_{10} N_{\text{HI}}$	$f(N_{\text{HI}}) (10^{-21})$	68% limits (10^{-21})	95% limits (10^{-21})
20.0 – 20.1	3.288	3.240 – 3.338	3.192 – 3.386
20.1 – 20.2	0.715	0.690 – 0.741	0.667 – 0.766
20.2 – 20.3	0.381	0.366 – 0.399	0.351 – 0.414
20.3 – 20.4	0.261	0.249 – 0.273	0.238 – 0.284
20.4 – 20.5	0.192	0.184 – 0.202	0.175 – 0.211
20.5 – 20.6	0.123	0.117 – 0.130	0.111 – 0.136
20.6 – 20.7	8.83×10^{-2}	$[8.35 - 9.31] \times 10^{-2}$	$[7.93 - 9.78] \times 10^{-2}$
20.7 – 20.8	6.42×10^{-2}	$[6.07 - 6.78] \times 10^{-2}$	$[5.73 - 7.13] \times 10^{-2}$
20.8 – 20.9	4.29×10^{-2}	$[4.05 - 4.55] \times 10^{-2}$	$[3.82 - 4.80] \times 10^{-2}$
20.9 – 21.0	2.47×10^{-2}	$[2.30 - 2.65] \times 10^{-2}$	$[2.15 - 2.81] \times 10^{-2}$
21.0 – 21.1	1.73×10^{-2}	$[1.63 - 1.86] \times 10^{-2}$	$[1.53 - 1.97] \times 10^{-2}$
21.1 – 21.2	9.24×10^{-3}	$[8.58 - 10.09] \times 10^{-3}$	$[7.83 - 10.85] \times 10^{-3}$
21.2 – 21.3	6.89×10^{-3}	$[6.44 - 7.49] \times 10^{-3}$	$[5.92 - 8.02] \times 10^{-3}$
21.3 – 21.4	3.75×10^{-3}	$[3.45 - 4.11] \times 10^{-3}$	$[3.15 - 4.46] \times 10^{-3}$
21.4 – 21.5	2.03×10^{-3}	$[1.84 - 2.32] \times 10^{-3}$	$[1.65 - 2.51] \times 10^{-3}$
21.5 – 21.6	1.35×10^{-3}	$[1.20 - 1.54] \times 10^{-3}$	$[1.09 - 1.69] \times 10^{-3}$
21.6 – 21.7	8.05×10^{-4}	$[7.16 - 9.25] \times 10^{-4}$	$[6.26 - 10.44] \times 10^{-4}$
21.7 – 21.8	3.79×10^{-4}	$[3.08 - 4.74] \times 10^{-4}$	$[2.61 - 5.45] \times 10^{-4}$
21.8 – 21.9	2.45×10^{-4}	$[2.07 - 3.01] \times 10^{-4}$	$[1.69 - 3.58] \times 10^{-4}$
21.9 – 22.0	7.48×10^{-5}	$[5.98 - 11.96] \times 10^{-5}$	$[2.99 - 14.95] \times 10^{-5}$
22.0 – 22.1	2.38×10^{-5}	$[1.19 - 4.75] \times 10^{-5}$	$0 - 7.13 \times 10^{-5}$
22.1 – 22.2	2.83×10^{-5}	$[1.89 - 4.72] \times 10^{-5}$	$[9.43 - 66.03] \times 10^{-6}$
22.2 – 22.3	2.25×10^{-5}	$[1.50 - 3.75] \times 10^{-5}$	$[7.49 - 44.96] \times 10^{-6}$
22.3 – 22.4	5.95×10^{-6}	$[5.95 - 17.86] \times 10^{-6}$	$0 - 2.98 \times 10^{-5}$
22.4 – 22.5	4.73×10^{-6}	$0 - 1.42 \times 10^{-5}$	$0 - 1.89 \times 10^{-5}$
22.5 – 22.6	3.76×10^{-6}	$0 - 1.13 \times 10^{-5}$	$0 - 1.50 \times 10^{-5}$
22.6 – 22.7	2.98×10^{-6}	$0 - 8.95 \times 10^{-6}$	$0 - 1.19 \times 10^{-5}$
22.7 – 22.8	0	$0 - 4.74 \times 10^{-6}$	$0 - 7.11 \times 10^{-6}$
22.8 – 22.9	0	$0 - 1.88 \times 10^{-6}$	$0 - 3.76 \times 10^{-6}$
22.9 – 23.0	0	$0 - 1.50 \times 10^{-6}$	$0 - 1.50 \times 10^{-6}$

Table A4. Column density distribution function for all DLAs with $3 < z < 4$. See also Figure 2.

$\log_{10} N_{\text{HI}}$	$f(N_{\text{HI}})$ (10^{-21})	68% limits (10^{-21})	95% limits (10^{-21})
20.0 – 20.1	2.735	2.547 – 2.962	2.377 – 3.150
20.1 – 20.2	0.839	0.734 – 0.959	0.629 – 1.079
20.2 – 20.3	0.500	0.428 – 0.595	0.357 – 0.667
20.3 – 20.4	0.284	0.236 – 0.350	0.189 – 0.397
20.4 – 20.5	0.203	0.165 – 0.248	0.128 – 0.285
20.5 – 20.6	0.155	0.125 – 0.191	0.101 – 0.221
20.6 – 20.7	0.118	$[9.48 - 14.69] \times 10^{-2}$	$[7.58 - 17.06] \times 10^{-2}$
20.7 – 20.8	7.15×10^{-2}	$[5.65 - 9.41] \times 10^{-2}$	$[4.14 - 10.92] \times 10^{-2}$
20.8 – 20.9	5.68×10^{-2}	$[4.48 - 7.18] \times 10^{-2}$	$[3.59 - 8.67] \times 10^{-2}$
20.9 – 21.0	3.80×10^{-2}	$[3.09 - 4.99] \times 10^{-2}$	$[2.14 - 5.94] \times 10^{-2}$
21.0 – 21.1	2.83×10^{-2}	$[2.26 - 3.77] \times 10^{-2}$	$[1.70 - 4.34] \times 10^{-2}$
21.1 – 21.2	1.65×10^{-2}	$[1.20 - 2.25] \times 10^{-2}$	$[8.99 - 26.97] \times 10^{-3}$
21.2 – 21.3	1.07×10^{-2}	$[8.33 - 15.47] \times 10^{-3}$	$[4.76 - 19.04] \times 10^{-3}$
21.3 – 21.4	7.56×10^{-3}	$[4.73 - 10.40] \times 10^{-3}$	$[2.84 - 13.24] \times 10^{-3}$
21.4 – 21.5	3.76×10^{-3}	$[2.25 - 5.26] \times 10^{-3}$	$[1.50 - 6.76] \times 10^{-3}$
21.5 – 21.6	2.39×10^{-3}	$[1.19 - 3.58] \times 10^{-3}$	$[5.97 - 47.72] \times 10^{-4}$
21.6 – 21.7	9.48×10^{-4}	$[4.74 - 18.95] \times 10^{-4}$	$0 - 2.37 \times 10^{-3}$
21.7 – 21.8	0	$0 - 7.53 \times 10^{-4}$	$0 - 1.13 \times 10^{-3}$
21.8 – 21.9	2.99×10^{-4}	$0 - 8.97 \times 10^{-4}$	$0 - 1.20 \times 10^{-3}$
21.9 – 22.0	2.37×10^{-4}	$0 - 7.12 \times 10^{-4}$	$0 - 9.50 \times 10^{-4}$
22.0 – 22.1	3.77×10^{-4}	$[3.77 - 5.66] \times 10^{-4}$	$[1.89 - 7.55] \times 10^{-4}$
22.1 – 22.2	0	$0 - 3.00 \times 10^{-4}$	$0 - 3.00 \times 10^{-4}$
22.2 – 22.3	0	$0 - 1.19 \times 10^{-4}$	$0 - 2.38 \times 10^{-4}$
22.3 – 22.4	0	$0 - 1.89 \times 10^{-4}$	$0 - 1.89 \times 10^{-4}$
22.4 – 22.5	0	$0 - 1.50 \times 10^{-4}$	$0 - 1.50 \times 10^{-4}$
22.5 – 22.6	0	$0 - 5.97 \times 10^{-5}$	$0 - 1.19 \times 10^{-4}$
22.6 – 22.7	0	$0 - 4.74 \times 10^{-5}$	$0 - 9.48 \times 10^{-5}$
22.7 – 22.8	0	$0 - 3.76 \times 10^{-5}$	$0 - 7.53 \times 10^{-5}$
22.8 – 22.9	0	$0 - 2.99 \times 10^{-5}$	$0 - 2.99 \times 10^{-5}$

Table A5. Column density distribution function for all DLAs with $4 < z < 5$. See also Figure 2.

z	dN/dX	68% limits	95% limits	$\Omega_{\text{DLA}}(10^{-3})$	68% limits	95% limits
2.00 – 2.17	0.0540	0.0531 – 0.0550	0.0521 – 0.0559	0.783	0.738 – 0.834	0.698 – 0.887
2.17 – 2.33	0.0584	0.0575 – 0.0594	0.0565 – 0.0604	0.710	0.680 – 0.745	0.654 – 0.784
2.33 – 2.50	0.0584	0.0573 – 0.0595	0.0562 – 0.0606	0.744	0.706 – 0.789	0.676 – 0.836
2.50 – 2.67	0.0594	0.0582 – 0.0607	0.0569 – 0.0620	0.700	0.672 – 0.733	0.648 – 0.769
2.67 – 2.83	0.0712	0.0696 – 0.0728	0.0681 – 0.0744	0.784	0.755 – 0.816	0.729 – 0.849
2.83 – 3.00	0.0763	0.0745 – 0.0783	0.0727 – 0.0801	0.815	0.787 – 0.846	0.762 – 0.878
3.00 – 3.17	0.0813	0.0791 – 0.0836	0.0770 – 0.0858	0.946	0.910 – 0.985	0.878 – 1.024
3.17 – 3.33	0.0831	0.0803 – 0.0860	0.0776 – 0.0888	0.954	0.908 – 1.003	0.866 – 1.054
3.33 – 3.50	0.0903	0.0868 – 0.0943	0.0832 – 0.0979	1.094	1.001 – 1.211	0.925 – 1.343
3.50 – 3.67	0.0840	0.0797 – 0.0885	0.0757 – 0.0925	1.258	1.095 – 1.473	0.973 – 1.728
3.67 – 3.83	0.0742	0.0695 – 0.0798	0.0648 – 0.0849	1.104	0.846 – 1.391	0.694 – 1.679
3.83 – 4.00	0.0730	0.0662 – 0.0806	0.0601 – 0.0875	0.861	0.741 – 1.016	0.642 – 1.329
4.00 – 4.17	0.0796	0.0706 – 0.0899	0.0616 – 0.0988	1.108	0.906 – 1.329	0.730 – 1.563
4.17 – 4.33	0.1033	0.0894 – 0.1192	0.0775 – 0.1331	1.275	1.126 – 1.439	0.991 – 1.619
4.33 – 4.50	0.1203	0.1027 – 0.1408	0.0880 – 0.1584	2.335	1.609 – 3.608	1.250 – 5.201
4.50 – 4.67	0.1620	0.1358 – 0.1927	0.1139 – 0.2190	2.609	2.158 – 3.126	1.735 – 3.697
4.67 – 4.83	0.1881	0.1580 – 0.2333	0.1204 – 0.2709	1.920	1.510 – 2.370	1.150 – 2.839
4.83 – 5.00	0.0956	0.0638 – 0.1435	0.0319 – 0.1913	1.340	0.696 – 2.125	0.221 – 3.014

Table A6. Table of dN/dX and Ω_{DLA} from our catalogue for $2 < z < 5$. See also Figure 3 and Figure 4.